A comprehensive analysis of natural sequence variation within the Arabidopsis thaliana nuclear auxin signaling pathway using an accessible web application

2018-05-24

Table of Contents

knitr::opts\_chunk$set(warning = FALSE, collapse = TRUE, comment = "#>", tidy = TRUE, cache = TRUE, out.width = "90%", fig.height=4, fig.width=7, out.extra='style="margin: auto; display: block; padding-top: 15px;"')  
  
#devtools::install\_github("wrightrc/r1001genomes", ref = "auxin-natural-variation")  
library(r1001genomes)

## Loading required package: plyr

## Loading required package: DECIPHER

## Loading required package: Biostrings

## Loading required package: BiocGenerics

## Loading required package: parallel

##   
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':  
##   
## clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,  
## clusterExport, clusterMap, parApply, parCapply, parLapply,  
## parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':  
##   
## IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':  
##   
## anyDuplicated, append, as.data.frame, cbind, colMeans,  
## colnames, colSums, do.call, duplicated, eval, evalq, Filter,  
## Find, get, grep, grepl, intersect, is.unsorted, lapply,  
## lengths, Map, mapply, match, mget, order, paste, pmax,  
## pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce,  
## rowMeans, rownames, rowSums, sapply, setdiff, sort, table,  
## tapply, union, unique, unsplit, which, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##   
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:plyr':  
##   
## rename

## The following object is masked from 'package:base':  
##   
## expand.grid

## Loading required package: IRanges

##   
## Attaching package: 'IRanges'

## The following object is masked from 'package:plyr':  
##   
## desc

## Loading required package: XVector

##   
## Attaching package: 'XVector'

## The following object is masked from 'package:plyr':  
##   
## compact

##   
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':  
##   
## strsplit

## Loading required package: RSQLite

## Warning: replacing previous import 'Biostrings::collapse' by  
## 'dplyr::collapse' when loading 'r1001genomes'

## Warning: replacing previous import 'Biostrings::intersect' by  
## 'dplyr::intersect' when loading 'r1001genomes'

## Warning: replacing previous import 'biomaRt::select' by 'dplyr::select'  
## when loading 'r1001genomes'

## Warning: replacing previous import 'Biostrings::setdiff' by  
## 'dplyr::setdiff' when loading 'r1001genomes'

## Warning: replacing previous import 'Biostrings::setequal' by  
## 'dplyr::setequal' when loading 'r1001genomes'

## Warning: replacing previous import 'Biostrings::union' by 'dplyr::union'  
## when loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::failwith' by 'plyr::failwith'  
## when loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::id' by 'plyr::id' when loading  
## 'r1001genomes'

## Warning: replacing previous import 'dplyr::summarize' by 'plyr::summarize'  
## when loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::count' by 'plyr::count' when  
## loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::desc' by 'plyr::desc' when  
## loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::mutate' by 'plyr::mutate' when  
## loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::arrange' by 'plyr::arrange' when  
## loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::rename' by 'plyr::rename' when  
## loading 'r1001genomes'

## Warning: replacing previous import 'dplyr::summarise' by 'plyr::summarise'  
## when loading 'r1001genomes'

library(DT)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 2.2.1 ✔ purrr 0.2.4  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.4  
## ✔ tidyr 0.8.0 ✔ stringr 1.3.0  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::arrange() masks plyr::arrange()  
## ✖ dplyr::collapse() masks Biostrings::collapse(), IRanges::collapse()  
## ✖ dplyr::combine() masks BiocGenerics::combine()  
## ✖ purrr::compact() masks XVector::compact(), plyr::compact()  
## ✖ dplyr::count() masks plyr::count()  
## ✖ dplyr::desc() masks IRanges::desc(), plyr::desc()  
## ✖ tidyr::expand() masks S4Vectors::expand()  
## ✖ dplyr::failwith() masks plyr::failwith()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::first() masks S4Vectors::first()  
## ✖ dplyr::id() masks plyr::id()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::mutate() masks plyr::mutate()  
## ✖ ggplot2::Position() masks BiocGenerics::Position(), base::Position()  
## ✖ purrr::reduce() masks IRanges::reduce()  
## ✖ dplyr::rename() masks S4Vectors::rename(), plyr::rename()  
## ✖ dplyr::slice() masks XVector::slice(), IRanges::slice()  
## ✖ dplyr::summarise() masks plyr::summarise()  
## ✖ dplyr::summarize() masks plyr::summarize()

library(magrittr)

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

library(RColorBrewer)  
library(ggpmisc)

## For news about 'ggpmisc', please, see http://www.r4photobiology.info/

## For on-line documentation see http://docs.r4photobiology.info/ggpmisc/

library(ggthemes)  
library(ggseqlogo)  
library(ggtree)

## Loading required package: treeio

##   
## Attaching package: 'treeio'

## The following object is masked from 'package:Biostrings':  
##   
## mask

## ggtree v1.10.5 For help: https://guangchuangyu.github.io/ggtree  
##   
## If you use ggtree in published research, please cite:  
## Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 2017, 8(1):28-36, doi:10.1111/2041-210X.12628

##   
## Attaching package: 'ggtree'

## The following object is masked from 'package:magrittr':  
##   
## inset

## The following object is masked from 'package:dplyr':  
##   
## collapse

## The following object is masked from 'package:tidyr':  
##   
## expand

## The following object is masked from 'package:Biostrings':  
##   
## collapse

## The following object is masked from 'package:IRanges':  
##   
## collapse

## The following object is masked from 'package:S4Vectors':  
##   
## expand

## The following object is masked from 'package:plyr':  
##   
## .

library(ape)

## Warning: package 'ape' was built under R version 3.4.4

##   
## Attaching package: 'ape'

## The following object is masked from 'package:ggtree':  
##   
## rotate

## The following objects are masked from 'package:treeio':  
##   
## drop.tip, Nnode, Ntip

## The following object is masked from 'package:Biostrings':  
##   
## complement

library(reticulate)  
library(scales)

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

library(viridis)

## Warning: package 'viridis' was built under R version 3.4.4

## Loading required package: viridisLite

##   
## Attaching package: 'viridis'

## The following object is masked from 'package:scales':  
##   
## viridis\_pal

use\_python(python = "/usr/local/bin/python")

# List of working titles

* A comprehensive analysis of natural sequence variation within the Arabidopsis thaliana nuclear auxin signaling pathway
* **Add your suggestions**

# Abstract

* Motivating, high level statement
* Statement defining the cutting edge of your research area
* A statement of the form: “However, it remains unclear …” that defines what problem you are  
  addressing.
* A statement of the form:“Here, we …” that concisely states your contribution.
* A series of 3?5 statement about your results.
* A concluding statement.

# Intro

The first genome sequence of *Arabidopsis thaliana* facilitated rapid advancement of plant biology through molecular genetics. Since this initial genome, massive scaling of sequencing technology has allowed a pioneering group to survey the global genomic variation in natural *Arabidopsis thaliana* populations. This valuable population genetics resource has led to several associations of genetic loci with phenotypic traits. Outside of genome-wide studies this valuable dataset has seen little use. Natural genomic variation also provides a set of feasible functional variation at the gene, gene family and gene network level that may provide insight into the function and evolution of genes, families and networks. Here, we present a web application and R-package through which plant molecular biologists with little to no bioinformatics experience can make use of this rich dataset of genetic variation to formulate hypotheses as to the sequence/function/phenotype relationships determined by the gene, family, or network of their interest. We demonstrate the utility of this tool through comprehensive analysis and identification of potential functional variation in the nuclear auxin signaling pathway.

This application allows easy access to 1. lists of missense polymorphisms to facilitate biochemical assays of variant effects (Starita et al., [2017](#ref-starita_variant_2017)), 2. family-wise alignments of variants to facilitate de novo functional domain identification (Melamed, Young, Miller, & Fields, [2015](#ref-melamed_combining_2015)), 3. lists of accessions containing missense (or any type of) polymorphisms to facilitate segregation analysis and measurment of the distribution of variant effects on phenotype (Park et al., [2017](#ref-park_distributions_2017)), 4. as well as the study of gene and network evolution (**???**; Delker et al., [2010](#ref-delker_natural_2010)).

Here, we have utilized this tool to analyze the genetic variation in the nuclear auxin signaling pathway and formulate hypotheses regarding the functional implications of this variation and the evolution of the genes and gene families in this pathway. Our analysis also provided further confirmation of much of the existing knowledge of these genes demonstrating the validity of this approach.

# Methods

## Nucleotide Diversity Calculation

Nei defined the nucleotide diversity statistic in his original paper as: “…the average number of nucleotide differences per site between two randomly chosen DNA sequences.” (Nei, 1979) And provided the equation:

Where Where is the frequency of the ith sequence in the population and is the number of sites that are different between the ith and jth sequence divided by sequence length.

a more general form, that treats each sequence in the population as unique can be written as:

where is the nucleotide (A, T, C or G) at position k on the ith sequence of the population. is the length of the sequence (the same for all sequences in the population, assumes only SNPs exist, no indels). n is the total number of sequences in the population.

from this form we can re-arrange summations to the form below:

here can be thought of as the site-wise nucleotide diversity at position k, and is equal to the nucleotide diversity of a sequence of length 1 at location k. we can calculate for each site, then average those over the sequence length to calculate , the nucleotide diversity of the sequence.

The function Nucleotide\_diversity in the r1001genomes package calculates for each position in the gene or region that contains a variant. note, is equal to 0 at all locations without variants. This is also what is displayed in the Diversity Plot tab of the web app.

## Detailed calculaiton simplification.

The formula for above requires comparing every sequence to every other sequence at location k, however, we know there are only a few variant forms at each individual location.

so we can revert back to using Nei’s original formula, modifying it slightly, replacing with , being the number of sequences in the population of the form at location k: subscripts i and k are summed only over unique forms [need to clarify this]

we will define as the number of sequences different from i at k.

we can see the summed term will be zero if , and if so:

next we substitute our definition of

then distribute and split summation:

finally, summing is equal to n:

this simplified form for is used by the app, because the counts of unique variants at a single nucleotide location can easily be summarized in R.

# Results

## TIR1/AFB genes

Auxin acts by binding to receptors that in turn target co-corepressors for degradation. Auxin receptors (six in the model plant *Arabidopsis thaliana*) evolved through gene duplication and diversification early in the history of vascular plants (**???**). The rate of co-repressor degradation is determined by the identity of both the receptor and co-repressor (**???**), and that this rate sets the pace of development (**???**).

Although it is unknown whether different receptors play different roles at the organismal level, a number of studies have pointed out differences in the ways the six different receptors in A. thaliana differ in their expression. The expression of AFB5 is induced in the presence of auxin, making this a circadian-regulated gene (**???**). Also, AFB3 is more highly expressed in the roots in the presence of nitrate, allowing increased lateral root formation (**???**). Although TIR1/AFBs are expressed ubiquitously in A. thaliana tissues, TIR1, AFB2, and AFB3 have been shown to accumulate in the shoot and root meristems and leaf tissues, with slightly different expression patterns for TIR1 (**???**). AFB1 has a mutation drastically reducing its ability to assemble into an SCF complex, contributing to its increased accumulation, as well as preventing it from degrading Aux/IAAs(**???**). Higher order TIR1/AFB mutants show AFB1 to have inconclusive function. However, AFB1 is shown to be able to accumulate ubiquitously in all tissues (**???**). AFB4 and AFB5 have been shown to preferentially and functionally bind the synthetic auxin picloram, while others have a preference for naturally occurring auxin, indole-3-acetic acid (**???**). Interestingly, the strength and rate with which TIR1/AFBs are able to bind and mark Aux/IAAs for degradation are variable (**???**). TIR1 and AFB2 have slightly different speed at which they perform this function, with AFB2 being generally faster.

gene\_info <- geneInfoFromFile(fname = system.file("shiny-app", "gene\_family\_data",   
 "AFB\_gene\_ids.csv", package = "r1001genomes"))  
#> [1] "new genes:"  
#> character(0)  
annotation <- readAnnotationFile(filename = system.file("shiny-app", "gene\_family\_data",   
 "AFB\_annotations.csv", package = "r1001genomes"), gene\_info = gene\_info)  
alignment <- alignCDS(gene\_info$transcript\_ID)  
#> Determining distance matrix based on shared 5-mers:  
#>   
 |   
 | | 0%  
 |   
 |================== | 28%  
 |   
 |================================== | 52%  
 |   
 |============================================== | 71%  
 |   
 |======================================================= | 85%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Clustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |================== | 28%  
 |   
 |================================== | 52%  
 |   
 |============================================== | 71%  
 |   
 |======================================================= | 85%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Aligning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |========== | 16%  
 |   
 |===================== | 33%  
 |   
 |================================ | 50%  
 |   
 |=========================================== | 66%  
 |   
 |====================================================== | 83%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.27 secs  
#>   
#> Iteration 1 of 2:  
#>   
#> Determining distance matrix based on alignment:  
#>   
 |   
 | | 0%  
 |   
 |================== | 28%  
 |   
 |================================== | 52%  
 |   
 |============================================== | 71%  
 |   
 |======================================================= | 85%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Reclustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |================== | 28%  
 |   
 |================================== | 52%  
 |   
 |============================================== | 71%  
 |   
 |======================================================= | 85%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Realigning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Alignment converged - skipping remaining iteration.  
# make an alignment data frame  
aln\_df <- makeAlnDF(alignment[[2]])  
  
# make a collapsed alignment for translating positions  
collapsed\_aln\_df <- aln\_df %>% dplyr::select(aln\_pos, seq\_pos, seq\_name) %>%   
 tidyr::spread(key = seq\_name, value = seq\_pos)  
  
VCF <- readRDS(system.file("shiny-app", "gene\_family\_data", "AFBs\_VCF.rds",   
 package = "r1001genomes"))  
div\_stats <- ldply(VCF, diversityStats, geneInfo = gene\_info, .id = "transcript\_ID")  
formatRound(datatable(div\_stats, colnames = c("transcript", "symbol", "&pi;<sub>N</sub>",   
 "&pi;<sub>S</sub>", "&pi;<sub>N</sub>/&pi;<sub>S</sub>", "&pi; coding",   
 "&pi; transcript"), rownames = FALSE, escape = FALSE, options = list(paging = FALSE,   
 searching = FALSE)), columns = 2:7, digits = 6)

Examining the natural sequence variation across the AFB family revealed that *TIR1* and *AFB1* both had very low nonsynonymous diversity, hinting at their likely functional importance. *AFB3* and *AFB4* had higher nonsynonymous diversity, while their sisters genes, *AFB2* and *AFB5* were more conserved. This matches our current understanding of *AFB3* and *AFB4* as being minor players in the auxin signaling pathway, but brings in to question the inconclusive role of *AFB1* in auxin signaling.

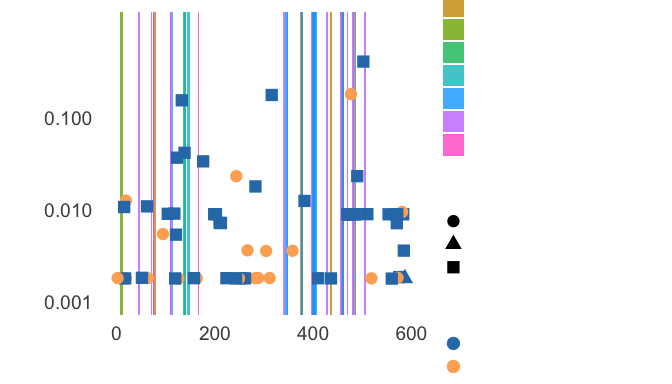
Although most known functional regions are highly conserved in AFB1, there is a nonsynonymous polymorphism in the TIR1/AFB oligomerization domain, only found in the Can-1 accession. Mutations in this domain of TIR1 frequently have a semidominant effect on root phenotypes (Dezfulian et al., [2016](#ref-dezfulian_oligomerization_2016); Wright, Zahler, Gerben, & Nemhauser, [2017](#ref-wright_insights_2017a)). Characterization of this accession may identify potential *AFB1* phenotypes and may also test the hypothesis that AFB1 functions through oligomerization with other TIR1/AFBs.

coding\_vcf <- ldply(.data = VCF, .fun = subset, !is.na(Transcript\_ID) & gt\_GT !=   
 "0|0")  
aln\_df <- addSNPsToAlnDF(aln\_df, coding\_vcf, seq\_name = Transcript\_ID, seq\_pos = Codon\_Number)  
collapsed\_aln\_df %<>% left\_join(y = subset(aln\_df, seq\_name == gene\_info[gene\_info$tair\_symbol ==   
 "TIR1", "transcript\_ID"]))  
  
#' Select variants for importing into PyMOL  
#'  
#' @param gene\_info a gene info data frame  
#' @param aln\_df an alignment data frame with attached SNPs  
#' @param structure\_ref string, which tair\_symbol or transcript\_ID (depending on   
#' \code{name\_by\_symbols}) corresponds to the protein structure to which   
#' selections should be mapped  
#' @param effect\_type string of the effect type to select from the alignment   
#' data frame  
#' @param name\_by\_symbols logical, if true (default) tair\_symbols will be used   
#' to name the selections, otherwise transcript\_IDs will be used  
#'  
#' @return  
#' @export  
#'  
#' @examples  
select\_variants <- function(gene\_info, aln\_df, structure\_ref, effect\_type = "missense",   
 name\_by\_symbols = TRUE) {  
 collapsed\_aln\_df <- aln\_df %>% dplyr::select(aln\_pos, seq\_pos, seq\_name) %>%   
 tidyr::spread(key = seq\_name, value = seq\_pos)  
 if (name\_by\_symbols)   
 column <- "tair\_symbol" else column <- "transcript\_ID"  
   
 selections <- ldply(.data = gene\_info[, column], .fun = function(gene) {  
 collapsed\_aln\_df %<>% dplyr::left\_join(y = subset(aln\_df, seq\_name ==   
 gene\_info[gene\_info[, column] == gene, "transcript\_ID"]))  
 variants <- collapsed\_aln\_df[grepl(effect\_type, collapsed\_aln\_df$effects),   
 gene\_info[gene\_info[, column] == structure\_ref, "transcript\_ID"]]  
 c(gene, paste0(na.omit(as.numeric(variants[[1]])), collapse = "+"))  
 })  
 return(selections)  
}  
  
selections <- select\_variants(gene\_info, aln\_df, structure\_ref = "TIR1")  
selections$V2 <- paste0("chain B and resi ", selections$V2)  
write.table(selections, file = "AFB\_variant\_selections", row.names = FALSE,   
 col.names = FALSE, sep = ",", quote = FALSE)

#  
# -- readSelection.py  
#  
from pymol import cmd, stored  
def readSelection (fileName):  
 '''  
 DESCRIPTION  
 reads a csv file containing a column of selection names and a column of  
 selection strings and creates selection objects for each  
 '''  
 #read in the selection file  
 with open(fileName) as f:  
 lines = f.readlines()  
 for line in lines:  
 sel = line.split(",")  
 cmd.select(sel[0], sel[1])  
 f.close()  
cmd.extend('readSelection',readSelection);

run ~/Desktop/natural-variation-webtool/manuscript/readSelection.py  
  
readSelection ~/Desktop/natural-variation-webtool/manuscript/AFB\_variant\_selections

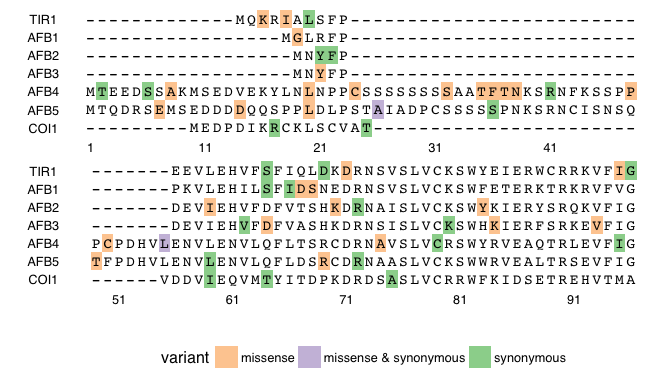
coding\_variants <- getCodingDiv(VCF[[gene\_info[gene\_info$tair\_symbol == "AFB1",   
 "transcript\_ID"]]])  
p <- plotCodingDiv(uniqueCodingVars = coding\_variants)  
annotation <- addAlnPosToAnno(anno\_df = annotation, aln\_df)  
annotation  
#> $domains  
#> tair\_symbol annotation annotation\_type position\_type reference\_PMID  
#> 1 TIR1 Fbox domain AA 17410169  
#> 2 TIR1 LRR domain AA 17410169  
#> 3 TIR1 miR393 miRNA AA 20018756  
#> 4 COI1 Fbox domain AA 17410169  
#> 5 COI1 LRR domain AA 17410169  
#> 6 COI1 miR393 miRNA AA 20018756  
#> 7 AFB5 Fbox domain AA 17410169  
#> 8 AFB5 LRR domain AA 17410169  
#> 9 AFB5 miR393 miRNA AA 20018756  
#> 10 AFB4 Fbox domain AA 17410169  
#> 11 AFB4 LRR domain AA 17410169  
#> 12 AFB4 miR393 miRNA AA 20018756  
#> 13 AFB3 Fbox domain AA 17410169  
#> 14 AFB3 LRR domain AA 17410169  
#> 15 AFB3 miR393 miRNA AA 20018756  
#> 16 AFB2 Fbox domain AA 17410169  
#> 17 AFB2 LRR domain AA 17410169  
#> 18 AFB2 miR393 miRNA AA 20018756  
#> 19 AFB1 Fbox domain AA 17410169  
#> 20 AFB1 LRR domain AA 17410169  
#> 21 AFB1 miR393 miRNA AA 20018756  
#> tair\_locus transcript\_ID chromosome\_name start\_position end\_position  
#> 1 AT3G62980 AT3G62980.1 3 23273065 23276630  
#> 2 AT3G62980 AT3G62980.1 3 23273065 23276630  
#> 3 AT3G62980 AT3G62980.1 3 23273065 23276630  
#> 4 AT2G39940 AT2G39940.1 2 16672493 16675864  
#> 5 AT2G39940 AT2G39940.1 2 16672493 16675864  
#> 6 AT2G39940 AT2G39940.1 2 16672493 16675864  
#> 7 AT5G49980 AT5G49980.1 5 20333897 20336671  
#> 8 AT5G49980 AT5G49980.1 5 20333897 20336671  
#> 9 AT5G49980 AT5G49980.1 5 20333897 20336671  
#> 10 AT4G24390 AT4G24390.1 4 12613076 12616238  
#> 11 AT4G24390 AT4G24390.1 4 12613076 12616238  
#> 12 AT4G24390 AT4G24390.1 4 12613076 12616238  
#> 13 AT1G12820 AT1G12820.1 1 4368760 4371298  
#> 14 AT1G12820 AT1G12820.1 1 4368760 4371298  
#> 15 AT1G12820 AT1G12820.1 1 4368760 4371298  
#> 16 AT3G26810 AT3G26810.1 3 9867652 9870782  
#> 17 AT3G26810 AT3G26810.1 3 9867652 9870782  
#> 18 AT3G26810 AT3G26810.1 3 9867652 9870782  
#> 19 AT4G03190 AT4G03190.1 4 1404887 1407402  
#> 20 AT4G03190 AT4G03190.1 4 1404887 1407402  
#> 21 AT4G03190 AT4G03190.1 4 1404887 1407402  
#> strand transcript\_start transcript\_end regionString  
#> 1 -1 23273479 23276181 3:23273479-23276181  
#> 2 -1 23273479 23276181 3:23273479-23276181  
#> 3 -1 23273479 23276181 3:23273479-23276181  
#> 4 -1 16672848 16675486 2:16672848-16675486  
#> 5 -1 16672848 16675486 2:16672848-16675486  
#> 6 -1 16672848 16675486 2:16672848-16675486  
#> 7 -1 20334420 20336531 5:20334420-20336531  
#> 8 -1 20334420 20336531 5:20334420-20336531  
#> 9 -1 20334420 20336531 5:20334420-20336531  
#> 10 -1 12613909 12615966 4:12613909-12615966  
#> 11 -1 12613909 12615966 4:12613909-12615966  
#> 12 -1 12613909 12615966 4:12613909-12615966  
#> 13 -1 4368879 4370780 1:4368879-4370780  
#> 14 -1 4368879 4370780 1:4368879-4370780  
#> 15 -1 4368879 4370780 1:4368879-4370780  
#> 16 1 9868342 9870464 3:9868342-9870464  
#> 17 1 9868342 9870464 3:9868342-9870464  
#> 18 1 9868342 9870464 3:9868342-9870464  
#> 19 -1 1405108 1407057 4:1405108-1407057  
#> 20 -1 1405108 1407057 4:1405108-1407057  
#> 21 -1 1405108 1407057 4:1405108-1407057  
#> transcript\_length end start seq\_name.x end\_aln\_pos seq\_name.y  
#> 1 2702 51 1 AT3G62980.1 96 AT3G62980.1  
#> 2 2702 594 51 AT3G62980.1 655 AT3G62980.1  
#> 3 2702 512 506 AT3G62980.1 568 AT3G62980.1  
#> 4 2638 58 1 AT2G39940.1 96 AT2G39940.1  
#> 5 2638 592 58 AT2G39940.1 641 AT2G39940.1  
#> 6 2638 523 517 AT2G39940.1 572 AT2G39940.1  
#> 7 2111 96 50 AT5G49980.1 96 AT5G49980.1  
#> 8 2111 619 96 AT5G49980.1 637 AT5G49980.1  
#> 9 2111 558 552 AT5G49980.1 568 AT5G49980.1  
#> 10 2057 96 50 AT4G24390.1 96 AT4G24390.1  
#> 11 2057 623 96 AT4G24390.1 637 AT4G24390.1  
#> 12 2057 558 552 AT4G24390.1 568 AT4G24390.1  
#> 13 1901 46 1 AT1G12820.1 96 AT1G12820.1  
#> 14 1901 577 46 AT1G12820.1 637 AT1G12820.1  
#> 15 1901 509 503 AT1G12820.1 568 AT1G12820.1  
#> 16 2122 46 1 AT3G26810.1 96 AT3G26810.1  
#> 17 2122 575 46 AT3G26810.1 637 AT3G26810.1  
#> 18 2122 507 501 AT3G26810.1 568 AT3G26810.1  
#> 19 1949 47 1 AT4G03190.1 96 AT4G03190.1  
#> 20 1949 585 47 AT4G03190.1 651 AT4G03190.1  
#> 21 1949 508 502 AT4G03190.1 568 AT4G03190.1  
#> start\_aln\_pos  
#> 1 14  
#> 2 96  
#> 3 562  
#> 4 10  
#> 5 96  
#> 6 566  
#> 7 50  
#> 8 96  
#> 9 562  
#> 10 50  
#> 11 96  
#> 12 562  
#> 13 19  
#> 14 96  
#> 15 562  
#> 16 19  
#> 17 96  
#> 18 562  
#> 19 18  
#> 20 96  
#> 21 562  
#>   
#> $positions  
#> tair\_symbol annotation annotation\_type position position\_type  
#> 1 TIR1 auxiaa\_contact binding 82 AA  
#> 2 TIR1 auxin\_contact binding 78 AA  
#> 3 TIR1 auxin\_contact binding 79 AA  
#> 4 TIR1 auxin\_contact binding 378 AA  
#> 5 TIR1 auxin\_contact binding 380 AA  
#> 6 TIR1 auxin\_contact binding 403 AA  
#> 7 TIR1 auxin\_contact binding 405 AA  
#> 8 TIR1 auxin\_contact binding 438 AA  
#> 9 TIR1 auxin\_contact binding 440 AA  
#> 10 TIR1 auxin\_contact binding 462 AA  
#> 11 TIR1 auxin\_contact binding 464 AA  
#> 12 TIR1 auxin\_contact binding 487 AA  
#> 13 TIR1 CUL1\_int binding 12 AA  
#> 14 TIR1 CUL1\_int binding 15 AA  
#> 15 TIR1 dimer\_int mutant 140 AA  
#> 16 TIR1 dimer\_null mutant 117 AA  
#> 17 TIR1 dimer\_null mutant 142 AA  
#> 18 TIR1 dimer\_null mutant 143 AA  
#> 19 TIR1 dimer\_null mutant 147 AA  
#> 20 TIR1 dimer\_null mutant 151 AA  
#> 21 TIR1 iaa\_contact binding 347 AA  
#> 22 TIR1 iaa\_contact binding 350 AA  
#> 23 TIR1 iaa\_contact binding 351 AA  
#> 24 TIR1 iaa\_contact binding 380 AA  
#> 25 TIR1 iaa\_contact binding 405 AA  
#> 26 TIR1 iaa\_contact binding 407 AA  
#> 27 TIR1 iaa\_contact binding 409 AA  
#> 28 TIR1 iaa\_contact binding 464 AA  
#> 29 TIR1 iaa\_contact binding 465 AA  
#> 30 TIR1 iaa\_contact binding 489 AA  
#> 31 TIR1 insP\_contacts binding 49 AA  
#> 32 TIR1 insP\_contacts binding 74 AA  
#> 33 TIR1 insP\_contacts binding 78 AA  
#> 34 TIR1 insP\_contacts binding 113 AA  
#> 35 TIR1 insP\_contacts binding 114 AA  
#> 36 TIR1 insP\_contacts binding 344 AA  
#> 37 TIR1 insP\_contacts binding 401 AA  
#> 38 TIR1 insP\_contacts binding 402 AA  
#> 39 TIR1 insP\_contacts binding 431 AA  
#> 40 TIR1 insP\_contacts binding 460 AA  
#> 41 TIR1 insP\_contacts binding 484 AA  
#> 42 TIR1 insP\_contacts binding 485 AA  
#> 43 TIR1 insP\_contacts binding 509 AA  
#> 44 TIR1 iaa\_contact binding 403 AA  
#> 45 TIR1 TIR1\_dm mutant 170 AA  
#> 46 TIR1 TIR1\_dm mutant 473 AA  
#> reference\_PMID tair\_locus transcript\_ID chromosome\_name start\_position  
#> 1 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 2 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 3 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 4 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 5 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 6 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 7 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 8 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 9 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 10 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 11 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 12 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 13 23539280 AT3G62980 AT3G62980.1 3 23273065  
#> 14 23539280 AT3G62980 AT3G62980.1 3 23273065  
#> 15 23539280 AT3G62980 AT3G62980.1 3 23273065  
#> 16 27618443 AT3G62980 AT3G62980.1 3 23273065  
#> 17 27618443 AT3G62980 AT3G62980.1 3 23273065  
#> 18 27618443 AT3G62980 AT3G62980.1 3 23273065  
#> 19 27618443 AT3G62980 AT3G62980.1 3 23273065  
#> 20 27618443 AT3G62980 AT3G62980.1 3 23273065  
#> 21 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 22 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 23 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 24 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 25 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 26 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 27 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 28 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 29 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 30 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 31 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 32 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 33 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 34 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 35 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 36 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 37 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 38 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 39 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 40 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 41 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 42 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 43 17410169 AT3G62980 AT3G62980.1 3 23273065  
#> 44 23539280 AT3G62980 AT3G62980.1 3 23273065  
#> 45 23539280 AT3G62980 AT3G62980.1 3 23273065  
#> 46 23539280 AT3G62980 AT3G62980.1 3 23273065  
#> end\_position strand transcript\_start transcript\_end regionString  
#> 1 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 2 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 3 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 4 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 5 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 6 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 7 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 8 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 9 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 10 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 11 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 12 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 13 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 14 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 15 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 16 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 17 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 18 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 19 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 20 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 21 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 22 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 23 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 24 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 25 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 26 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 27 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 28 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 29 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 30 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 31 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 32 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 33 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 34 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 35 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 36 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 37 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 38 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 39 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 40 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 41 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 42 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 43 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 44 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 45 23276630 -1 23273479 23276181 3:23273479-23276181  
#> 46 23276630 -1 23273479 23276181 3:23273479-23276181  
#> transcript\_length seq\_name aln\_pos  
#> 1 2702 AT3G62980.1 127  
#> 2 2702 AT3G62980.1 123  
#> 3 2702 AT3G62980.1 124  
#> 4 2702 AT3G62980.1 431  
#> 5 2702 AT3G62980.1 433  
#> 6 2702 AT3G62980.1 456  
#> 7 2702 AT3G62980.1 458  
#> 8 2702 AT3G62980.1 491  
#> 9 2702 AT3G62980.1 496  
#> 10 2702 AT3G62980.1 518  
#> 11 2702 AT3G62980.1 520  
#> 12 2702 AT3G62980.1 543  
#> 13 2702 AT3G62980.1 57  
#> 14 2702 AT3G62980.1 60  
#> 15 2702 AT3G62980.1 186  
#> 16 2702 AT3G62980.1 162  
#> 17 2702 AT3G62980.1 188  
#> 18 2702 AT3G62980.1 189  
#> 19 2702 AT3G62980.1 193  
#> 20 2702 AT3G62980.1 197  
#> 21 2702 AT3G62980.1 398  
#> 22 2702 AT3G62980.1 401  
#> 23 2702 AT3G62980.1 402  
#> 24 2702 AT3G62980.1 433  
#> 25 2702 AT3G62980.1 458  
#> 26 2702 AT3G62980.1 460  
#> 27 2702 AT3G62980.1 462  
#> 28 2702 AT3G62980.1 520  
#> 29 2702 AT3G62980.1 521  
#> 30 2702 AT3G62980.1 545  
#> 31 2702 AT3G62980.1 94  
#> 32 2702 AT3G62980.1 119  
#> 33 2702 AT3G62980.1 123  
#> 34 2702 AT3G62980.1 158  
#> 35 2702 AT3G62980.1 159  
#> 36 2702 AT3G62980.1 395  
#> 37 2702 AT3G62980.1 454  
#> 38 2702 AT3G62980.1 455  
#> 39 2702 AT3G62980.1 484  
#> 40 2702 AT3G62980.1 516  
#> 41 2702 AT3G62980.1 540  
#> 42 2702 AT3G62980.1 541  
#> 43 2702 AT3G62980.1 565  
#> 44 2702 AT3G62980.1 456  
#> 45 2702 AT3G62980.1 216  
#> 46 2702 AT3G62980.1 529  
# translate positions from TIR1 to AFB1  
  
AFB1\_positions <- annotation$positions  
AFB1\_positions %<>% left\_join(y = select(collapsed\_aln\_df, aln\_pos, gene\_info[gene\_info$tair\_symbol ==   
 "AFB1", "transcript\_ID"]), by = "aln\_pos")  
AFB1\_positions$position <- AFB1\_positions$AT4G03190.1  
  
p <- append\_layers(p, list(geom\_rect(data = AFB1\_positions, mapping = aes(xmin = as.integer(position) -   
 1, xmax = as.integer(position) + 2, fill = annotation), ymin = -Inf, ymax = Inf,   
 inherit.aes = FALSE, alpha = 0.8)), position = "bottom")  
p



The AFB4 and AFB5 receptors have an N-terminal extension prior to the F-box domains. This extension had very high nonsynonymous diversity(Fig. 4), suggesting that this extension does not play an important functional role in these proteins.

Figure 4: Alignment of A. thaliana TIR1/AFB F-box domains showing variants. Cyan marks the F-box binding domain. Yellow, green, and magenta, mark missense and synonymous, synonymous only, and missense only variant locations respectively. Yellow squared E’s show Cul1-interacting residues, differing in AFB1.

aln\_df <- makeAlnDF(alignment[[2]])  
aln\_df <- left\_join(aln\_df, dplyr::select(gene\_info, "tair\_locus", "tair\_symbol",   
 "transcript\_ID"), by = c(transcript\_ID = "transcript\_ID"))  
aln\_df$tair\_symbol <- fct\_rev(fct\_relevel(aln\_df$tair\_symbol, "TIR1"))  
aln\_df <- chunkAlnDF(aln\_df = aln\_df, chunk\_width = 55)  
VCF <- ldply(.data = VCF, .fun = subset, gt\_GT != "0|0")  
VCF <- getCodingDiv(VCF)  
aln\_df <- addSNPsToAlnDF(aln\_df, VCF)  
F\_box <- subset(aln\_df, aln\_pos <= 96)  
F\_box <- chunkAlnDF(aln\_df = F\_box, chunk\_width = 80)  
p <- ggplot(F\_box, aes(x = aln\_pos, y = tair\_symbol, group = seq\_pos, text = variants))  
p <- p + geom\_tile(data = na.omit(F\_box), mapping = aes(fill = effects), width = 1,   
 height = 1, alpha = 0.8) + geom\_text(aes(label = letter), alpha = 1, family = "Courier") +   
 scale\_fill\_brewer(type = "qual", palette = 1, direction = -1) + scale\_x\_continuous(breaks = seq(1,   
 max(F\_box$aln\_pos), by = 10)) + scale\_y\_discrete() + # expand increases distance from axis  
labs(x = "", y = "", fill = "variant") + theme\_logo(base\_family = "Helvetica") +   
 theme(panel.grid = element\_blank(), panel.grid.minor = element\_blank()) +   
 facet\_wrap(facets = ~chunk, ncol = 1, scales = "free") + theme(strip.background = element\_blank(),   
 strip.text.x = element\_blank(), legend.box = "vertical")  
p



## Aux/IAA genes

The Aux/IAA proteins (IAAs) mediate both auxin perception and transcriptional co-repression of auxin responsive genes. Auxin perception is carried out along with TIR1/AFB proteins, which together with IAAs act as a co-receptor to bind and perceive intracellular auxin. Interaction of IAAs with auxin and TIR1/AFBs is conferred by the highly conserved 13 amino acid degron motif (Tan et al., [2007](#ref-tan_mechanism_2007)). Formation of this auxin co-receptor complex triggers ubiquitination and subsequent degradation of IAAs. This degradation in turn relieves their repression upon ARF transcription factors.Stabilizing mutations within the degron can lead to dominant phenotypes (Berleth, Krogan, & Scarpella, [2004](#ref-berleth_auxin_2004); Yang et al., [2004](#ref-yang_iaa1_2004)). The *Arabidopsis thaliana* Aux/IAA gene family has 29 members, most of which possess a canonical degron. Notable exceptions are the auxin-insensitive IAA20 and IAA30 that lack a degron sequence entirely, while the closely related IAA31 retains a semi-functional degron sequence with a very slow degradation rate (Dreher, Brown, Saw, & Callis, [2006](#ref-dreher_arabidopsis_2006); Havens et al., [2012](#ref-havens_synthetic_2012)). Additionally, IAA degradation rate is influenced by rate motifs found outside the degron, such as the KR motif (Dreher et al., [2006](#ref-dreher_arabidopsis_2006); Moss et al., [2015](#ref-moss_rate_2015)). IAA-mediated repression of ARFs depends upon both an EAR domain, N-terminal of the degron, that recruits co-repressors of the TOPLESS (TPL) family, and a Phox and Bem1 (PB1) domain that enables interaction via hetero-oligomerization with ARF PB1 domains (Kagale & Rozwadowski, [2011](#ref-kagale_ear_2011); Nanao et al., [2014](#ref-nanao_structural_2014)).

gene\_info <- geneInfoFromFile(fname = system.file("shiny-app", "gene\_family\_data",   
 "IAA\_gene\_ids.csv", package = "r1001genomes"))  
#> [1] "new genes:"  
#> character(0)  
# annotation <- readAnnotationFile(filename = system.file('shiny-app',  
# 'gene\_family\_data', 'IAA\_annotations.csv', package = 'r1001genomes'),  
# gene\_info = gene\_info)  
alignment <- alignCDS(gene\_info$transcript\_ID)  
#> Determining distance matrix based on shared 4-mers:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Clustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.01 secs  
#>   
#> Aligning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |== | 3%  
 |   
 |===== | 7%  
 |   
 |====== | 10%  
 |   
 |========= | 14%  
 |   
 |=========== | 17%  
 |   
 |============== | 21%  
 |   
 |================ | 25%  
 |   
 |================== | 28%  
 |   
 |===================== | 32%  
 |   
 |======================= | 35%  
 |   
 |========================= | 39%  
 |   
 |=========================== | 42%  
 |   
 |============================== | 46%  
 |   
 |================================ | 50%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================= | 60%  
 |   
 |========================================== | 64%  
 |   
 |============================================ | 67%  
 |   
 |============================================== | 71%  
 |   
 |================================================= | 75%  
 |   
 |=================================================== | 78%  
 |   
 |===================================================== | 82%  
 |   
 |======================================================= | 85%  
 |   
 |========================================================== | 89%  
 |   
 |============================================================ | 92%  
 |   
 |============================================================== | 96%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.37 secs  
#>   
#> Iteration 1 of 2:  
#>   
#> Determining distance matrix based on alignment:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Reclustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.02 secs  
#>   
#> Realigning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |=== | 5%  
 |   
 |====== | 10%  
 |   
 |========== | 15%  
 |   
 |============= | 20%  
 |   
 |================ | 25%  
 |   
 |==================== | 30%  
 |   
 |======================= | 35%  
 |   
 |========================== | 40%  
 |   
 |============================= | 45%  
 |   
 |================================ | 50%  
 |   
 |==================================== | 55%  
 |   
 |======================================= | 60%  
 |   
 |========================================== | 65%  
 |   
 |============================================== | 70%  
 |   
 |================================================= | 75%  
 |   
 |==================================================== | 80%  
 |   
 |======================================================= | 85%  
 |   
 |========================================================== | 90%  
 |   
 |============================================================== | 95%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.38 secs  
#>   
#> Iteration 2 of 2:  
#>   
#> Determining distance matrix based on alignment:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0 secs  
#>   
#> Reclustering into groups by similarity:  
#>   
 |   
 | | 0%  
 |   
 |==== | 6%  
 |   
 |======== | 13%  
 |   
 |============ | 19%  
 |   
 |================= | 26%  
 |   
 |===================== | 32%  
 |   
 |======================== | 37%  
 |   
 |============================ | 43%  
 |   
 |=============================== | 48%  
 |   
 |================================== | 53%  
 |   
 |===================================== | 57%  
 |   
 |======================================== | 62%  
 |   
 |=========================================== | 66%  
 |   
 |============================================== | 70%  
 |   
 |================================================ | 74%  
 |   
 |================================================== | 77%  
 |   
 |==================================================== | 80%  
 |   
 |====================================================== | 83%  
 |   
 |======================================================== | 86%  
 |   
 |========================================================= | 88%  
 |   
 |=========================================================== | 91%  
 |   
 |============================================================ | 93%  
 |   
 |============================================================= | 94%  
 |   
 |============================================================== | 96%  
 |   
 |=============================================================== | 97%  
 |   
 |================================================================ | 98%  
 |   
 |================================================================ | 99%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.02 secs  
#>   
#> Realigning Sequences:  
#>   
 |   
 | | 0%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.02 secs  
#>   
#> Refining the alignment:  
#>   
 |   
 | | 0%  
 |   
 |======= | 11%  
 |   
 |============================= | 44%  
 |   
 |==================================== | 56%  
 |   
 |=========================================== | 67%  
 |   
 |=================================================== | 78%  
 |   
 |=================================================================| 100%  
#>   
#> Time difference of 0.15 secs  
# make an alignment data frame  
aln\_df <- makeAlnDF(alignment[[2]])  
  
# make a collapsed alignment for translating positions  
collapsed\_aln\_df <- aln\_df %>% dplyr::select(aln\_pos, seq\_pos, seq\_name) %>%   
 tidyr::spread(key = seq\_name, value = seq\_pos)  
  
VCF <- readRDS(system.file("shiny-app", "gene\_family\_data", "IAAs\_VCF.rds",   
 package = "r1001genomes"))  
div\_stats <- ldply(VCF, diversityStats, geneInfo = gene\_info, .id = "transcript\_ID")  
div\_stats$tair\_symbol %<>% fct\_reorder(.x = , div\_stats$tair\_symbol, .fun = function(x) {  
 as.integer(str\_remove(x, pattern = "IAA"))  
})  
formatRound(datatable(div\_stats[order(div\_stats$tair\_symbol), ], colnames = c("transcript",   
 "symbol", "&pi;<sub>N</sub>", "&pi;<sub>S</sub>", "&pi;<sub>N</sub>/&pi;<sub>S</sub>",   
 "&pi; coding", "&pi; transcript"), rownames = FALSE, escape = FALSE, options = list(paging = FALSE,   
 searching = FALSE)), columns = 2:7, digits = 6)

Within the available natural sequence variation, we found that for the canonical Aux/IAAs, no missense mutations were present in the KR motif or residues responsible for PB1-PB1 domain interactions (Supp. Fig. Alignment of canonical IAAs). The majority of nonsynonymous mutations observed were in the variable regions between the conserved domains. However, several missense polymorphisms of interest were identified in EAR domains and degrons of several canonical IAAs.

### Sister Pair Diversification

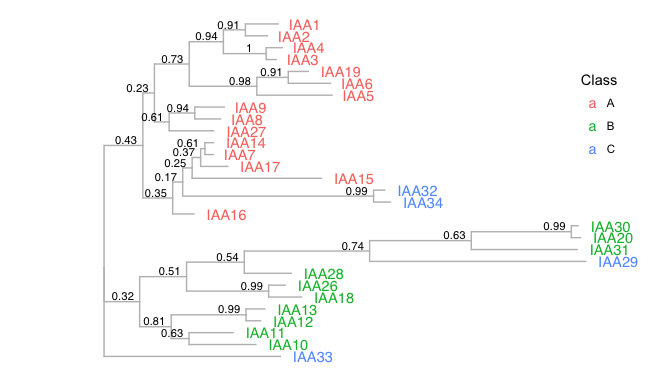
Previous research has ~~pointed to~~ **found evidence of** both broad genetic redundancy and specificity within closely related pairs or groups of IAA proteins (Overvoorde et al., [2005](#ref-overvoorde_functional_2005)). For example, the iaa8-1, iaa9-1 double mutant and the iaa5-1, iaa6-1, iaa19-1 triple mutant have wild-type phenotypes (Overvoorde et al., [2005](#ref-overvoorde_functional_2005)), yet the IAA6/19 sister pair has significant differences in expression patterns, protein abundances and functions, suggesting they have undergone functional specialization since their divergence (Winkler et al., [2017](#ref-winkler_variation_2017)).

In our analysis of the full IAA gene family, one member of most IAA sister pairs had high nonsynonymous diversity while the other sister was more conserved (Figure IAA phylogeny heat map). In an interspecific comparison between *A. thaliana* and *A. lyrata*, IAA19 is more conserved, while IAA6 has regions of increased sequence divergence upstream of the degron and in the N-terminal half of the PB1 domain (Winkler et al., [2017](#ref-winkler_variation_2017)). Consistent with these results, our analysis of intraspecific natural variation revealed higher conservation for IAA19 ( = 0.55) compared to IAA6 ( = 2.3) (Figure IAA phylogeny heat map), and also detected high diversity within the same regions of IAA6 as seen in Winkler et al. (Supp Fig IAA6 diversity plot).

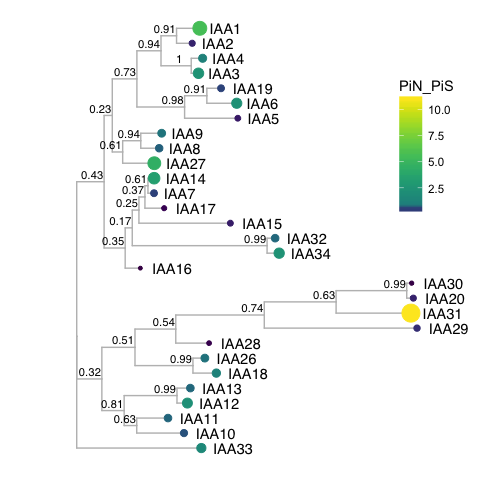
write.nexus.data(as.matrix(alignment[[2]]), file = "IAAs.nex", format = "protein")  
aliscore(as.AAbin(alignment[[2]]), path = "/Volumes/SSD/Applications/ALISCORE\_v2.0/Aliscore\_v.2.0/")  
id <- scan(file = "input.fas\_List\_random.txt", sep = " ", quiet = TRUE)  
AAAlign\_no\_gaps <- as.matrix(alignment[[2]])[, id]  
write.nexus.data(AAAlign\_no\_gaps, file = "manuscript/IAAs.nex", format = "protein")

mb  
  
execute manuscript/IAAs.nex  
# set IAA33 as the outgroup  
outgroup AT5G57420.1  
  
lset nucmodel=protein Rates=invgamma  
# or include Nst=6 ngammacat=4 omegavar=M3   
prset aamodelpr=fixed(jones) statefreqpr=fixed(empirical)   
# initial run with aamodelpr=mixed yielded posterior prob of 1 for jones model  
propset ExtTBR$prob=0  
  
mcmcp ngen=250000 nchains=4 printfreq=1000 samplefreq=100 burnin=200  
  
mcmc  
sump  
sumt

tree <- treeio::read.beast(file = "IAAs.nex.con.tre")  
  
  
IAA\_data <- read.csv("IAAs\_table.csv")  
# tip\_labels need to be in the first column  
IAA\_data %<>% select(name, everything())  
IAA\_data$name %<>% as.character()  
IAA\_data$AGI %<>% as.character()  
  
IAA\_data <- IAA\_data[match(gsub(x = tree@phylo$tip.label, pattern = "\\.1",   
 replacement = ""), IAA\_data$AGI), ]  
tree@phylo$tip.label <- IAA\_data$name  
ggtree(tree, layout = "rectangular", color = "grey") %<+% IAA\_data + theme\_tree() +   
 scale\_x\_continuous(expand = c(0, 1.2)) + geom\_tiplab(mapping = aes(color = Class),   
 hjust = -0.25) + theme(legend.position = c(0.9, 0.7)) + geom\_nodelab(mapping = aes(label = round(prob,   
 digits = 2)), nudge\_x = -0.3, nudge\_y = 0.5, size = 3)



p <- ggtree(tree, layout = "rectangular", color = "grey") %<+% IAA\_data + theme\_tree() +   
 scale\_x\_continuous(expand = c(0, 1.2)) + geom\_tiplab(hjust = -0.25) + geom\_tippoint(mapping = aes(size = PiN\_PiS,   
 color = PiN\_PiS), show.legend = TRUE) + scale\_color\_viridis(values = rescale(c(min(IAA\_data$PiN\_PiS),   
 1, max(IAA\_data$PiN\_PiS)), to = c(0, 1)), guide = "colourbar") + scale\_size(guide = "none") +   
 theme(legend.position = c(0.9, 0.7)) + geom\_nodelab(mapping = aes(label = round(prob,   
 digits = 2)), nudge\_x = -0.3, nudge\_y = 0.5, size = 3)  
p



Within the Group B IAA genes, there are four sister pairs—IAA10/IAA11, IAA12/IAA13, IAA18/IAA26, and IAA20/IAA30 (Fig X) (Remington, Vision, Guilfoyle, & Reed, [2004](#ref-remington_contrasting_2004)). Interestingly, among these sister pairs only IAA12 and 13 have a noticeable difference in nonsynonymous diversity. The *Arabidopsis thaliana* EFP browser shows that IAA10 and 11 have almost identical expression patterns, with the exception of IAA10 being more highly expressed in the developing seed, and IAA11 being more highly expressed in root tissue during lateral root development. In our analysis, IAA10 and IAA11 both showed functional conservation ( of 0.80 and 0.67 respectively) in support of IAA10 and IAA11 playing redundant roles. Accessions and associated variants identified here provide a potentially expedited route to elucidating the extent of redundancy between family members such as IAA10 and 11.

### Identification of new Loss-of-Function/Gain-of-Function mutations

Accessions with functional variants can be potentially used to investigate the extent of redundancy/specialization within sister pairs. For example, it has been shown that IAA10 has a missense mutation in the EAR domain, which converts the second leucine in the LxLxL motif to a serine (L49S), a mutation likely to reduce the ability of the IAA to recruit the TOPLESS co-repressor (M.-S. Lee, An, & Cho, [2016](#ref-lee_biological_2016)). This mutation was found in three accessions from the southern UK. A second potentially stronger mutation, represented in one accession from northeastern Spain is a ~~missense~~ **frameshift?** resulting in an early stop codon in the PB1 domain (need Supp. Fig?). This likely prevents IAA10 from interacting with an ARF. ~~These alleles would be strong candidates for loss-of-function mutants, allowing researchers to discover the function of IAA10. In particular, it would make possible a straightforward path for creating an iaa10/11 double mutant using CRISPR/Cas9 directed mutagenesis, allowing for examination of the extent of redundancy between these family members.~~ **I don’t think we can say this. There are SALK lines in the 5’ UTR and first intron of IAA10 and IAA11.**

Several examples of possible gain-of-function degron missense variants were found in IAA2, IAA3, IAA6, IAA7, IAA8, IAA13, IAA15, IAA18, IAA19, IAA27, IAA28, IAA29. Most degron variants were only found in one or two accessions (Supplemental Table or Map?). ~Interestingly, almost all of the IAAs found to have accessions with degron missense mutations have a closely related sister, which may indicate a redundant function within the sister-pairs.~ **But we know that iaa6 iaa19 iaa5 triple doesn’t have a phenotype.** IAA6 has two potentially dominant missense mutations found within the degron (Supp Fig. alignment), one previously believed to lead to the shy1-1 phenotype (Kim, Soh, Kang, Furuya, & Nam, [1996](#ref-kim_two_1996)). Interestingly, this variant is widely represented in 31 accessions across Europe (Supplemental Map Fig). This finding supports the proposition that iaa6[C78R] may not be the mutation causing the shy1-1 phenotype but rather represents natural variation in the degron sequence ([note from Mike Prigge in TAIR](https://www.arabidopsis.org/servlets/TairObject?type=notepad&id=11247)). These accessions could be used to confirm whether iaa6[C78R] is responsible for the SHY1-1D phenotype. Additionally, an accession was identified with the IAA6-V77M variant, also in the degron, that may be useful to further explore iaa6 gain-of-function phenotypes.

### Identifying novel functional sequences in IAAs

An exciting possible use for the Natural Variation Webtool is to search for conserved, low-diversity sites that may represent possible sites of post-translational modification or protein-protein interaction. For example, recent detailed biochemical analysis identified putative lysine ubiquitination sites in members of the IAA6/19 sister pairs (Winkler et al., [2017](#ref-winkler_variation_2017)); Supp Fig IAA6/19 alignment). These sites are conserved across natural sequence variation, while several other lysine residues vary (Supp Figs alignment and map). In IAA6, there are 17 lysines of which 6 are putative ubiquitination sites, one is in the KR motif, while two others show natural variation. In IAA19, there are 15 total lysines, with 8 as putative ubiquitination sites, 2 in the KR motif and degron, and 1 that exhibits natural variation. **This sentence is not clear. Only 1 of the ubiquitination sites varies?** Natural sequence diversity may be used to narrow down a subset of lysines to assess for functional ubiquitination.

The members of the IAA8/9/27 clade have a unique N-terminal extension with a highly conserved region common to orthologs in several other species (Dreher et al., [2006](#ref-dreher_arabidopsis_2006)). IAA8/9 play redundant roles in leaf serration (Koenig, Bayer, Kang, Kuhlemeier, & Sinha, [2009](#ref-koenig_auxin_2009)), while IAA8 important in lateral roots (Arase et al., [2012](#ref-arase_iaa8_2012)). IAA8 has been reported to uniquely localize in the cytosol (Arase et al., [2012](#ref-arase_iaa8_2012)) where it may interact with LSD1 protein (Coll, Epple, & Dangl, [2011](#ref-coll_programmed_2011)). The N-terminal extension has a conserved region shared between IAA8 and IAA9 with low natural sequence diversity (Supp Fig IAA8/9 alignment) that may represent a site of novel functionality, possibly mediating interaction with LSD1. There are a few accessions with missense alleles in this region in both IAA8 and IAA9 that may be valuable in probing the functionality of this region (Supplemental Fig IAA8/9 map).

Members of the IAA17/7/14/16 group have a second EAR motif (LxLxL) with no observed missense mutations in IAA7, 16, or 17 (Supp Fig IAA7/14/16/17 alignment). This observation supports the experimental evidence for IAA7 that the second EAR motif is functional: it can interact with TPR1 co-repressor and has a minor repressive role in auxin-related developmental phenotypes [lee\_biological\_2016]. For IAA14, we did observe two missense mutations that may be of interest to probe functionality of the IAA14 EAR motif: one in the central lysine (L41F) that is represented in only one accession, and the other in the second “x” (K40N) that is found in 12 accessions around Copenhagen (Supp Fig. IAA14 map).

## TPL/TPR genes

The Auxin signaling pathway utilizes the TOPLESS (TPL) and TOPLESS-related (TPR) family of Gro/TLE/TUP1 type co-repressor proteins to maintain auxin responsive genes in a transcriptionally-repressed state in the absence of auxin (cite). In *Arabidopsis thaliana* the TPL/TPR family includes *TPL* and four *TPR1-4*. The resulting proteins are comprised of three structural domains: an N-terminal TPL domain and two WD-40 domains. TPL/TPR proteins are recruited to the AUX/IAA proteins through interaction with the conserved Ethylene-responsive element binding factor-associated amphiphilic repression (EAR) domain (cite). Canonical EAR domains have the amino acid sequence LxLxL, as found in most AUX/IAAs (figure xA) (Overvoorde et al., [2005](#ref-overvoorde_functional_2005)). TPL/TPR co-repressors bind EAR domains via their C-terminal to LisH (CTLH) domains found near their N-termini (citations of pre-structure founding papers/reviews). Recent structural analyses of the TPL N-terminal domain have highlighted the precise interaction interface between TPL and AUX/IAA EAR domains, as well as the TPL-TPL dimerization and tetramerization motifs (Avarelilo et al 2017, Ke et al, 2015). The residues required for higher-order multimers of TPL tetramers have also been identified (Ma et al, 2017). Additional interactions with transcriptional regulation and chromatin modifying machinery are likely mediated by two tandem beta propeller domains of TPL/TPRs. These domain have not yet been crystallized, however numerous high resolution crystal structures of beta propeller domains allow for structural prediction of the TPL c-terminal beta propeller domains and the locations of likely interaction faces (Figure, X).

The TOPLESS co-repressor family generally exhibits a high level of sequence conservation at the amino acid sequence level across resequenced *Arabidopsis thaliana* accessions, with all values below 1 (Table X). The closely related TPL and TPR1 have the highest values (TPL-0.425, TPR1-0.739), suggesting that these these two related genes tolerate a higher degree of sequence and potentially functional diversity compared to TPR2-4. The N-terminal TPL domain of the TPL/TPR family is particularly conserved (Figure X). All non-synonymous polymorphisms observed in this region are either in the coils between helices or are highly conservative mutations within helices (i.e. Valine to Isoleucine), which would be predicted to exhibit little effect on TPL folding and function.

The high degree of conservation in the entire N-terminal domain underscores its importance. For example, the initial tpl-1 mutation (N176H) in the ninth helix is a dominant gain-of-function allele (Long et al, 2002, PMID12050130), which is capable of binding wild-type TPL protein and inducing protein aggregation (Ma, 2017, pmid28630893). It is therefore understandable that very little diversity is observed in the ninth alpha helix of TPL/TPRs as mutations in this domain could act in a dominant negative fashion, rendering any de novo mutation in this region embryo lethal. However, the entire N-terminal domain exhibits very low rates of nucleotide diversity, suggesting that each domain in this region is critical at the population level. Indeed, the crystal structure of TPL has revealed that the TPL domain is responsible for the multiple roles of TPL/TPR proteins: The LisH domain and CRA domain coordinate TPL dimerization and tetramerization respectively (26601214, 28698367), the CTLH domain is critical for interaction with DNA-binding transcription factors (26601214), and helix 9 within the CRA domain may be critical for the formation of higher-order TPL complexes (multimers of TPL tetramers, 28630893).

Compared to the TPL domain, both the proline rich and WD-40 domains exhibit a higher rate of missense mutations. We utilized the Phyre2.0 modeling software to predict the structure of the TPL/TPR WD-40 repeats in order to map non-synonymous mutations onto a visible protein chassis (Figure X). Here we can see both faces of each predicted beta-propeller structure of the WD-40 domains, which are connected by a short peptide linker. This model reveals amino acid conservation in the core and central residues of the WD-40 domains. ~~These observations could be used as the basis for forward mutagenesis strategies to test whether this region of TPL functions similarly to its homolog TLE, which binds its target genes EH1 and the C-terminal WRPW/Y motif (Hairy/Hes/Runx) in the center of the WD-40 domain (pmid 16762837).~~ **Can we say something more general here? There is not enough background to support this.**

Despite the presence of 5 TPL/TPR genes in the genome, which are considered to be redundant (pmid 16763149), very few SNPs are observed that are predicted to result in loss of function for any TPL/TPR gene. Only one frameshift variant was observed, TPR4-665delG, in a single accession (Aiell-4 from Italy). This deletion occurs in codon 222, just after the TPL N-terminal domain, and generates an early stop 11 codons downstream from the deletion. It is unclear whether this N-terminal fragment would be expressed at any appreciable level, as it is possible that the generation of an early stop codon would trigger nonsense-mediated decay. However, if expressed, it would be fascinating to examine whether this naturally occurring N-terminal truncation retains activity as a co-repressor. Given the observation that the TPL N-terminal domain is sufficient to bind histones (28630893), and repress transcription via synthetic assays in yeast (24979769), it is possible that this represents a semi-functional TPL variant. There has been no evidence for prioritization of individual TPL/TPR family members in the auxin response pathway, but experiments modulating the gene copy number of TPR1 revealed the importance of maintaining the proper dosage of TPL/TPR in order to accurately trigger the salicylic acid innate immune response (20647385). These results, taken together with the absence of TPL/TPR loss-of-function alleles in the *Arabidopsis thaliana* accessions suggest that perhaps the entire family is being conserved at the nucleotide level in the population to meet a gene dosage requirement, with little opportunity for neo- or sub-functionalization.

## ARF genes

Auxin response is ultimately mediated by the auxin responsive transcription factors (ARFs). There are 23 ARFs in *Arabidopsis thaliana* that are divided into three phylogenetic classes. Class A ARFs (ARF5, ARF6, ARF7, ARF8 and ARF19) activate transcription. These ARFs have a glutamine-rich region in the middle of the protein that may mediate activation (Guilefoyle, 2007). It has recently been shown that the middle region of ARF5 interacts with the SWI/SNF chromatin remodeling ATPases BRAMA and SPLAYED, possibly to reduce nucleosome occupancy and allow for the recruitment of transcription machinery (Wu, 2016). Additionally, ARF7 interacts with Mediator subunits, directly tethering transcriptional activation machinery to its binding sites in the chromosome (Ito, 2016). Class B and C ARFs are historically categorized as repressor ARFs, though the mechanism through which they confer repression has not been identified. Their middle regions tend to be proline- and serine-rich (citation).

Canonical ARFs are comprised of three major domains. Recent crystallization of these domain have informed structure-function analysis of the ARFs (Boer, 2014; Nanao, 2014; Korasick, 2014). These domains are conserved throughout land plants. ARFs share an N-terminal B3 DNA binding domain. Flanking this DNA-binding domain is a dimerization domain, which folds up into a single “taco-shaped” domain to allow for dimerization between ARFs. There is an auxiliary domain that immediately follows and interacts with the dimerization domain. The middle region is the most variable between ARFs, as mentioned above, but is characterized by repetitive units of glutamine (class A), serine, or proline residues (classes B and C).

The C-terminal domain of the ARF is a protein-protein interaction domain mediating interactions among ARFs, between ARFs and other transcription factors, and between ARFs and the Aux/IAA repressors. This interaction domain was recently characterized as a Phox and Bem1 (PB1) domain, which is comprised of a positive and negative face with conserved basic and acidic residues, respectively. The dipolar nature of the PB1 domain may mediate multimerization by the pairwise interaction of these faces on different proteins as the ARF7 PB1 domain was crystallized as a multimer (**???**). However, it is unclear whether this multimerization occurs or plays a developmental significant role in vivo. Interfering with ARF dimerization in either the DNA-binding proximal dimerization domain or the PB1 domain decreases the ability of class A ARFs to activate transcription in a heterologous yeast system (Pierre-Jerome, 2016).

While domain architecture is broadly conserved among the ARFs, there are exceptional cases. Three ARFs do not contain a PB1 domain at all, ARF3, ARF13, and ARF17, and several more have lost the conserved acidic or basic residues in the PB1 domain, suggesting they may be reduced to a single interaction domain. Several ARFs additionally have an expanded conserved region within the DNA-binding domain, of unknown function. The majority of domain variation among ARFs occurs in the large B-class subfamily. The liverwort *Marchantia polymorpha* has a single representative ARF of each class (Flores-Sandoval, 2015). The expansion of these classes in flowering plants is the result of both whole genome and tandem duplication events (Remington, 2004). The growth of the ARF family may have allowed for the expansion of the quantity and complexity of loci regulated by the ARFs and subsequent expansion in their regulation of developmental processes.

Class A ARFs are the most well-studied ARF subfamily—the five family members all act as transcriptional activators and have well-characterized, distinct developmental targets. Overall the diversity of class A ARFs was generally low, especially compared to the class B and C ARFs, suggesting that class A ARFs are central to auxin signal transduction and plant development. Analysis of class A ARF nonsynonymous diversity suggests that the majority of these ARFs are highly functionally conserved, with values much lower than 1 with the exception of ARF19, with value of 1.8. Comparing diversity within sister pairs, there is a similar trade-off as seen in most IAA sister pairs, with one sister being highly conserved and the other more divergent. ARF19 and ARF8 are the more divergent class A ARFs, with values at least three time those of their sisters, ARF7 and ARF6 respectively. This may suggest that ARF6 and ARF7 serve more essential purposes in plant development.

For all class A ARFs, the middle region of the protein was the predominant high diversity region. The DNA-binding domain had very few, low-diversity missense mutations, as did the PB1 domain. Considering the necessity of their conserved functions, the low level of variation in these key DNA and protein-protein interaction domains is expected.

Class A ARFs all contain the canonical B3 DNA-binding domain, PB1 interaction domain and glutamine-rich middle region (Guilefoyle, 2007). Within the middle region, ARF7 had several expansions of polyglutamine sequences compared to the other class A ARFs. Polyglutamine regions are known to readily expand and contract throughout evolutionary time due to replication error, and variation in polyglutamine length can be acted on by natural selection and have phenotypic consequences (Press, 2014). A few missense variants in ARF5 highlighted potential functional variation in the class A ARFs. Two highly represented variants resulting in substitutions of small amino acids for the bulky phenylalanine were found on the positive face of the PB1 domain, very close to the conserved lysine residue and may disrupt interactions with other ARF or Aux/IAA PB1 domains. Interestingly, these variants were always found together in accessions Tu-B2-3, Ru-2, Haes-1, Rd-0, 627ME-13Y1, Le-0, Bu-0 and Bsch-0. Another highly represented polymorphism in ARF5 results in change in the 17th amino acid from a leucine to a methionine. This position is in the N-terminal extension of ARF5 which is of unknown function. It is possible that this methionine acts as an alternative start site in the accessions.

### Class B ARFs

Andrew

ARF23, which is truncated in its DNA-binding region, has a high value of 4.1, and has many high-diversity valued nonsense variants, as may be expected of a gene that is pseudogenizing.

ARF3, ARF13, and ARF17 all lack a PB1 C-terminal domain, which suggests they may be unable to interact with Aux/IAA repressors and may act as an auxin-insensitive mechanism of adding competitive inhibition to ARF-mediated transcription. These ARFs span the range of ratios relative to the other ARFs, ranging from 0.85 to 4.1. ARF13, with the highest of 4.1 has many nonsense variants of high diversity.

ARF14, ARF15, ARF16, ARF20, and ARF21 all do not have one or multiple of the conserved acidic residues in the PB1 domain, and themselves form a subclade of the B ARFs. Interestingly, no non-synonymous variation is seen in these amino acid residues of these ARFs, suggesting that the amino acids that are in the regions may still be important for the function of these proteins.

### Class C ARFs

Hardik (All figure cite below are in folder “/Dropbox/natural-variation-webtool/manuscript\_ARF\_C/FIGURES”)

Class C ARFs show low nucleotide diversity scores, all pi\_N/pi\_S values substantially lower than 1. ARF16 was the most conserved, whereas its clade members (ARF10, ARF17) had scores at least four times higher (Remington et al., [2004](#ref-remington_contrasting_2004)).

Structurally, all three members of Class C ARFs contain a canonical B3 DNA-binding domain, but only ARF10 and ARF16 contain a PB1 domain. The DNA binding domains exhibit overall low diversity and presence of largely synonymous changes for all the ARFs. Only the ARF10 DNA binding domain contains missense variants (S147T and A157S) (Fig. ARF10\_DBD Variants distribution). **Any insights into whether these may affect function? They are quite conservative.** Of the PB1 domain containing class C ARFs, ARF16 exhibits several missense variants which are sporadically distributed, in contrast to the conserved PB1 domain of ARF10 (Fig. Arf16\_PB\_variants distribution) .

Several members of class B (ARF2, 3 and 4) and C ARFs (ARF10, 16 and 17) are distinguished from other ARFs on the basis of their post transcriptional regulation by trans-acting small interfering RNAs (tasiRNAs) and microRNAs (miRNAs) (Allen, Xie, Gustafson, & Carrington, [2005](#ref-allen_micrornadirected_2005); Rhoades et al., [2002](#ref-rhoades_prediction_2002)). **((???) Arf 6 and 8 targeted by miRNA 167 - check for conservation).** The miRNA target of class C ARFs showed a high degree of conservation (Fig. ARF10\_16\_17\_Alignment). This is consistent with perturbation studies that has shown drastic developmental phenotype for class C ARFs (**???**; Mallory, Bartel, & Bartel, [2005](#ref-mallory_micrornadirected_2005); J.-W. Wang et al., [2005](#ref-wang_control_2005)). For the class B ARFs (2, 3 and 4) tasiRNA binding motif is also highly conserved suggesting strong regulatory constrains at the locus (Fig ARF2\_3\_4\_Alignment). DNA binding domain of ARF4 exhibits non-synonymous variation in 4 accession and would be interesting to carry out phenotypic characterize them for developmental growth defects (Fig ARF4\_DBD Variants distribution).

# Discussion

Allen, E., Xie, Z., Gustafson, A. M., & Carrington, J. C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, *121*(2), 207–221. <https://doi.org/10.1016/j.cell.2005.04.004>

Arase, F., Nishitani, H., Egusa, M., Nishimoto, N., Sakurai, S., Sakamoto, N., & Kaminaka, H. (2012). IAA8 involved in lateral root formation interacts with the TIR1 auxin receptor and ARF transcription factors in Arabidopsis. *PloS One*, *7*(8), e43414. <https://doi.org/10.1371/journal.pone.0043414>

Berleth, T., Krogan, N. T., & Scarpella, E. (2004). Auxin signals–turning genes on and turning cells around. *Current Opinion in Plant Biology*, *7*(5), 553–563. <https://doi.org/10.1016/j.pbi.2004.07.016>

Coll, N. S., Epple, P., & Dangl, J. L. (2011). Programmed cell death in the plant immune system. *Cell Death and Differentiation*, *18*(8), 1247–1256. <https://doi.org/10.1038/cdd.2011.37>

Delker, C., Pöschl, Y., Raschke, A., Ullrich, K., Ettingshausen, S., Hauptmann, V., … Quint, M. (2010). Natural Variation of Transcriptional Auxin Response Networks in Arabidopsis thaliana. *The Plant Cell*, *22*(7), 2184–2200. <https://doi.org/10.1105/tpc.110.073957>

Dezfulian, M. H., Jalili, E., Roberto, D. K. A., Moss, B. L., Khoo, K., Nemhauser, J. L., & Crosby, W. L. (2016). Oligomerization of SCF TIR1 Is Essential for Aux/IAA Degradation and Auxin Signaling in Arabidopsis. *PLOS Genet*, *12*(9), e1006301. <https://doi.org/10.1371/journal.pgen.1006301>

Dreher, K. A., Brown, J., Saw, R. E., & Callis, J. (2006). The Arabidopsis Aux/IAA Protein Family Has Diversified in Degradation and Auxin Responsiveness. *The Plant Cell*, *18*(3), 699–714. <https://doi.org/10.1105/tpc.105.039172>

Havens, K. A., Guseman, J. M., Jang, S. S., Pierre-Jerome, E., Bolten, N., Klavins, E., & Nemhauser, J. L. (2012). A synthetic approach reveals extensive tunability of auxin signaling. *Plant Physiol*, *160*(1), 135–42. <https://doi.org/10.1104/pp.112.202184>

Kagale, S., & Rozwadowski, K. (2011). EAR motif-mediated transcriptional repression in plants: An underlying mechanism for epigenetic regulation of gene expression. *Epigenetics*, *6*(2), 141–146.

Kim, B. C., Soh, M. S., Kang, B. J., Furuya, M., & Nam, H. G. (1996). Two dominant photomorphogenic mutations of Arabidopsis thaliana identified as suppressor mutations of hy2. *The Plant Journal*, *9*(4), 441–456. <https://doi.org/10.1046/j.1365-313X.1996.09040441.x>

Koenig, D., Bayer, E., Kang, J., Kuhlemeier, C., & Sinha, N. (2009). Auxin patterns Solanum lycopersicum leaf morphogenesis. *Development (Cambridge, England)*, *136*(17), 2997–3006. <https://doi.org/10.1242/dev.033811>

Lee, M.-S., An, J.-H., & Cho, H.-T. (2016). Biological and molecular functions of two EAR motifs of Arabidopsis IAA7. *Journal of Plant Biology*, *59*(1), 24–32. <https://doi.org/10.1007/s12374-016-0453-1>

Mallory, A. C., Bartel, D. P., & Bartel, B. (2005). MicroRNA-directed regulation of Arabidopsis AUXIN RESPONSE FACTOR17 is essential for proper development and modulates expression of early auxin response genes. *The Plant Cell*, *17*(5), 1360–1375. <https://doi.org/10.1105/tpc.105.031716>

Melamed, D., Young, D. L., Miller, C. R., & Fields, S. (2015). Combining Natural Sequence Variation with High Throughput Mutational Data to Reveal Protein Interaction Sites. *PLoS Genet*, *11*(2), e1004918. <https://doi.org/10.1371/journal.pgen.1004918>

Moss, B. L., Mao, H., Guseman, J. M., Hinds, T. R., Hellmuth, A., Kovenock, M., … Nemhauser, J. L. (2015). Rate Motifs Tune Auxin/Indole-3-Acetic Acid Degradation Dynamics. *Plant Physiology*, *169*(1), 803–813. <https://doi.org/10.1104/pp.15.00587>

Nanao, M. H., Vinos-Poyo, T., Brunoud, G., Thévenon, E., Mazzoleni, M., Mast, D., … Dumas, R. (2014). Structural basis for oligomerization of auxin transcriptional regulators. *Nature Communications*, *5*, 3617. <https://doi.org/10.1038/ncomms4617>

Overvoorde, P. J., Okushima, Y., Alonso, J. M., Chan, A., Chang, C., Ecker, J. R., … Theologis, A. (2005). Functional Genomic Analysis of the AUXIN/INDOLE-3-ACETIC ACID Gene Family Members in Arabidopsis thaliana. *The Plant Cell*, *17*(12), 3282–3300. <https://doi.org/10.1105/tpc.105.036723>

Park, B., Rutter, M. T., Fenster, C. B., Symonds, V. V., Ungerer, M. C., & Townsend, J. P. (2017). Distributions of Mutational Effects and the Estimation of Directional Selection in Divergent Lineages of Arabidopsis thaliana. *Genetics*, *206*(4), 2105–2117. <https://doi.org/10.1534/genetics.116.199190>

Remington, D. L., Vision, T. J., Guilfoyle, T. J., & Reed, J. W. (2004). Contrasting modes of diversification in the Aux/IAA and ARF gene families. *Plant Physiology*, *135*(3), 1738–1752. <https://doi.org/10.1104/pp.104.039669>

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., & Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell*, *110*(4), 513–520.

Starita, L. M., Ahituv, N., Dunham, M. J., Kitzman, J. O., Roth, F. P., Seelig, G., … Fowler, D. M. (2017). Variant Interpretation: Functional Assays to the Rescue. *The American Journal of Human Genetics*, *101*(3), 315–325. <https://doi.org/10.1016/j.ajhg.2017.07.014>

Tan, X., Calderon-Villalobos, L. I. A., Sharon, M., Zheng, C., Robinson, C. V., Estelle, M., & Zheng, N. (2007). Mechanism of auxin perception by the TIR1 ubiquitin ligase. *Nature*, *446*(7136), 640–5. <https://doi.org/10.1038/nature05731>

Wang, J.-W., Wang, L.-J., Mao, Y.-B., Cai, W.-J., Xue, H.-W., & Chen, X.-Y. (2005). Control of root cap formation by MicroRNA-targeted auxin response factors in Arabidopsis. *The Plant Cell*, *17*(8), 2204–2216. <https://doi.org/10.1105/tpc.105.033076>

Winkler, M., Niemeyer, M., Hellmuth, A., Janitza, P., Christ, G., Samodelov, S. L., … Calderón Villalobos, L. I. A. (2017). Variation in auxin sensing guides AUX/IAA transcriptional repressor ubiquitylation and destruction. *Nature Communications*, *8*, 15706. <https://doi.org/10.1038/ncomms15706>

Wright, R. C., Zahler, M. L., Gerben, S. R., & Nemhauser, J. L. (2017). Insights into the Evolution and Function of Auxin Signaling F-Box Proteins in Arabidopsis thaliana Through Synthetic Analysis of Natural Variants. *Genetics*, *207*(2), 583–591. <https://doi.org/10.1534/genetics.117.300092>

Yang, X., Lee, S., So, J.-H., Dharmasiri, S., Dharmasiri, N., Ge, L., … Estelle, M. (2004). The IAA1 protein is encoded by AXR5 and is a substrate of SCF(TIR1). *The Plant Journal: For Cell and Molecular Biology*, *40*(5), 772–782. <https://doi.org/10.1111/j.1365-313X.2004.02254.x>